DOCUMENT RESUME

ED 069 759 TM 002 214

AUTHOR	Levy, Lynn B.; Fritz, Kentner V.
TITLE	Status Report on the Computer Grading of Essays.
INSTITUTION	Wisconsin Univ., Madison. Counseling Center.
PUB DATE	Jun 72
NOTE	14p.
JOURNAL CIT	Counseling Center Reports, v5 n10 June 1972
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Composition Skills (Literary); *Computers; *Essays; *Evaluation Techniques; *Grading; High School Students; Literature Reviews; Technical Reports
IDENTIFIERS	*Page (Ellis)

ABSTRACT

Writings on the use of computers to grade English compositions are reviewed in this article. A background is given on the work of Ellis Page, whose approach was to quantify the "indicators" of good writing and relate these to human judgments. Endeavors to grade the content as well as the style of student papers are also discussed. (Author/RS)

Counseling Center Reports

The University of Wisconsin 415 West Gilman Street Madison, Wisconsin 53706

Editor: Richard W. Johnson

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 069759

TM 002 214

Volume 5, Number 10 June, 1972

STATUS REPORT ON THE COMPUTER GRADING OF ESSAYS

Lynn B. Levy and Kentner V. Fritz

Abstract

The use of computers to grade English composition dates back some six or seven years to the pioneering work of Ellis Page of the University of Connecticut. Most subsequent experiments in the area have adopted Page's basic approach of quantifying the "indicators" of good writing and relating these to human judgments. Recently, endeavors have been made to grade the content as well as the style of student papers. The following pages briefly review the writings in the field from 1966 to the present.

Previous University of Wisconsin
Counseling Center Reports

Volume 4, 1970-71

No. 1 Thrush, R. S. Annual Report

No. 2 Johnson, R. W. Note on Measurement of Sex and Age Differences on SVIB-M.

No. 3 Kirk, K. W., Ohvall, R. A., and Johnson, R. W. Vocational Interests of Pharmacy Students.

No. 4 Praamsma, S. W., and Fritz, K. V. User's Manual for Item Diagnostic Program (IMDIAG).

No. 5 Johnson, R. W. Congruence of SVIB-W and KOIS Interest Profiles.

No. 6 Nolting, E., and Leege, W. Privileged Communication -- Rights and Responsibilities of College Counselors under Wisconsin Law.

No. 7 Mendoza, G. Prediction of Academic and Clinical Performance of Physical Therapy Students.

Volume 5, 1971-72

No. 1 Thrush, R. S. Annual Report

No. 2 Cornish, R. D. Annotated Bibliography of MMPI Research Among College Populations: 1962-1970.

No. 3 Fritz, K. V., and Cornish, R. D. A User's Guide to Scoring and Improving Examinations Using the MERMAC Test Analysis and Questionnaire Package.

No. 4 Johnson, R. W., and Schwertfeger, M. Relationship Between the Basic Interest Scales and the Occupational and Nonoccupational Scales on the Strong Vocational Interest Blank for Men.

No. 5 Johnson, R. W. Contradictory Scores on the Strong Vocational Interest Blank.

No. 6 Johnson, R. W. Content Analysis of Strong Vocational Interest Blank for Men.

No. 7 Johnson, R. W., and Johansson, C. B. Moderating Effect of Basic Interests on Predictive Validity of SVIB Occupational Scales.

No. 8 Fritz, K. V., and Levy L. Introduction to Computer Managed Instruction and the Automated Instructional Management System.

No. 9 Bennett, M. K. University of Wisconsin-Madison Norms for the Miller Analogies Test.

STATUS REPORT ON THE COMPUTER GRADING OF ESSAYS

Lynn B. Levy and Kentner V. Fritz
Counseling Center
University of Wisconsin - Madison

The computer grading of essays has been regarded by its critics as the dehumanizing imposition of a mechanical standard on student papers. To evaluate this phenomenon, however, we need to have an overview of what the computer does when it "grades" essays--and that is to emulate the grading procedures used by human judges, or, more precisely, to examine papers for the attributes of good writing agreed upon by human graders. Perhaps the first important implication of computer grading of essays is the awareness called for on the part of the grader of precisely what he is responding to in an essay, an awareness of his own cognitive process.

The published work in the area dates back to 1966. In that year both Ellis Page's "The Imminence of Grading Essays by Computer," and Arthur Daigon's "Computer Grading of English Composition" appeared. Page published "Grading Essays by Computers: Progress Report" in 1967 and "The Use of the Computer in Analyzing Student Essays" in 1968. In 1969, Jack Hiller, Donald Marcotte, and Timothy Martin co-authored "Opinionation, Vagueness, and Specificity-Distinctions: Essay Traits Measured by Computer." The first doctoral dissertation on computer grading of essays appeared in 1971, Henry Slotnick's "An Examination of the Computer Grading of Essays." That same year, "Essay Grading by Computer: A Laboratory Phenomenon?" by Henry Slotnick and John Knapp was published. And finally, in April of 1972, Thomas Knapp presented

"Essay Topics and Modes and their Effects on Student Prose" at the annual meeting of the American Educational Research Association.

Ellis Page was the first to use the computer to evaluate essays. He broke down the grade assigned any essay into two components, content and style. The central problem was then to identify qualities of content and style which could be programmed for use by the computer. This he did by dividing the qualities of writing into intrinsic (what he called the trins) and approximate (the proxes) qualities. Intrinsic qualities are those things human judges might look for such as good vocabulary, well constructed sentences of varying lengths, a smooth-flowing style, etc. Computers cannot be programmed to look for intrinsic qualities, but they can look for approximate qualities of writing--those quantifiable qualities logically related to intrinsic qualities. For example, the number of uncommon words used in an essay, the number of different words used, and the mean number of times each different word is used are indicators (or proxes) of the intrinsic quality of good vocabulary. The programmer can know that his choice of approximate qualities is correct when the machine's grade, based on the presence or absence of such indicators in the composition, correlates significantly with the grade assigned by skilled human judges. Thus approximation refers to simulation of the human product, without any great concern about the way this product was produced. According to Page, "all computer simulation of human behavior appears to be product simulation rather than process simulation" (1967).

Page's study, called Project Essay Grade, was performed in early 1965. Essays written by students in grades 8 through 12 at the

University of Wisconsin High School in Madison were judged by at least four independent graders for certain aspects of style believed independent of content. These judgments of overall quality formed the trins. Hypotheses were then generated about the variables which might be associated with these judgments. If these variables were measurable by computer and feasible to program within the logistics of the study, they became the proxes of the study. These included: whether a title was present, average sentence length, number of paragraphs, subject-verb openings, length of essay in words, average word length, standard deviation of word length, standard deviation of sentence length, numbers of parentheses, apostrophes, commas, periods, underlined words, dashes, colons, semi-colons, quotation marks, connective words, exclamation points, question marks, prepositions, spelling errors, relative pronouns, subordinating conjunctions, common words on the Dale list, hyphens, slashes, etc. Computer programs were written to measure these proxes in the essays. The essays were then keypunched and fed into the computer which generated data about the proxes and "scored" the papers. These scores were then analyzed for their multivariate relationship to the human ratings, were weighted approximately, and were used to maximize the prediction of the expert human ratings. All this was done by use of a standard multiple regression procedure. The results of the regression analysis included the establishment of weights to be assigned to each indicator or prox. The weights and the indicators' measurements appear in a prediction equation. Based on a grader's past performance, which resulted in the establishment of weights for the various indicators, additional

papers can be examined by the computer, weighted approximately, and the grades the judge would have assigned, had he evaluated the papers, can be predicted.

The set of measures Page used produced computer assigned grades statistically indistinguishable from those assigned by expert judges. The overall accuracy of this strategy was startling. The computer also did a good job predicting human judgments for a second set of essays written by the same students. From his results, Page predicted that in the future the computer would actually correlate better with human judges than will other humans.

Wishing greater detail in his analysis, Page broadened his categories to include five principle traits (adopted partly from the work of Paul Diederick of the ETS, Princeton): ideas, organization, style, mechanics, and creativity. In the summer of 1966, he called together a group of 32 highly qualified English teachers from Connecticut schools to see how they would handle all these traits, but especially creativity, in the grading process. Each of 256 essays was rated on a five point scale for each of these traits by eight such expert judges working independently. In the analysis of the teacher ratings, the same thirty proxes were used to investigate each of the five trait ratings. It was found that creativity was least reliably judged by these human experts, and mechanics most reliably graded. The computer, on the other hand, was more reliable with such difficult variables as creativity and organization.

Page speaks to the problem of the verbal education of a computer in order to extend its use to the humanities. "The solution," he

writes, "will probably be not in trying to program all the linguistic responses to be made by the computer. Rather the solution may consist in programming only a certain set of quasi-psychological procedures, designed to enable the computer to learn on its own (i.e., to gain literary experience) by reading in and correctly processing a great amount of appropriate text, making use of automated dictionaries and other aids while doing so. We dream of producing, in other words, the well-read computer. Part of our success to date has occurred through allowing the computer itself, in the multiple regression program, to determine which analytic weightings are valuable" (1967).

In his article (1966), Arthur Daigon talks briefly about Project Essay Grade, pointing out that "At present, it is in measurements related to style, that is, the selection of words and their syntactic arrangement, that the computer can tell us most about good, bad, or indifferent writing." He sees the failings of the teacher in not being enough like a machine. The teacher cannot accurately and consistently respond to discernible elements of style. The machine fails in that it is not enough like a human, who can respond to meaning, to connotations, to figurative language. We must attempt to make computers able to respond to substantive ideas, to "content." The machine must be able to answer these questions: Did the student do what was assigned? Did he deal with appropriate subject matter? Did he supply details, expand a definition, relate cause and effect, or use a chronological approach if any of these organizational modes were part of, or appropriate to the assignment?

Hiller, Marcotte, and Martin (1969) selected three characteristics

of writing (opinionation-exaggeration, vagueness, and specificity-distinctions) for study on the assumption that single words or discrete phrases reliably cue the presence of such characteristics in essays and that such characteristics are related to essay quality. They were interested in the development of measures capable not merely of increasing our ability to simulate teacher assigned grades, but of providing useful feedback on student performance to both student and teacher. They expected opinionation-exaggeration (authoritarian attitudes) and vagueness (ambiguity, haziness) to correlate negatively with essay quality, and specificity-distinctions (examples and illustrations provided, use of concrete language) to correlate positively. A total of 130 "cue" words or phrases was placed into the computer opinionation-exaggeration dictionary, 60 into its vagueness dictionary, and 90 into its specificity-distinctions dictionary.

They used a set of 256 essays written by students at the University of Wisconsin High School in 1963 and already used as a part of Project Essay Grade. These essays were written in one class period, the topic being a common saying: "The best things in life are free." Students were asked to agree, disagree, or state why they thought the statement was neither true nor false. These essays were then searched by computer for "cues" and the measures thus obtained were correlated with the essay grades. As expected, opinionation and vagueness were found to correlate negatively with the grading criteria, and specificity positively.

The study Slotnick reports in his University of Illinois dissertation (1971) attacked three questions: "Can objective indicators be

used to predict the assessments of human judges?"; Can the procedures producing significant correlations be cross-validated?"; and "How are the objective indicators related logically and in practice to the procedure by which grades are assigned to papers?" A set of 476 essays defending the proposition that money spent on the space program should have been spent on domestic problems was collected from sophomores, juniors, and seniors at Urbana High School in Urbana, Illinois. Eight judges,four of them high school English teachers and the other four teaching assistants in a freshman rhetoric course at the University of Illinois, each graded all the papers in twelve waves over 1½ days on a scale of one through five (one being well above average and five, well below). A multivariate analysis of variance was performed using the judges' ratings for content and style as dependent measures to determine whether judges responded differently to papers written by students of varying year in school and sex, or papers appearing in different waves, and whether any differences existed in the mean ratings assigned by the high school and college judges. No significant difference was found to exist in grades assigned to papers written by males and females or between the mean grades assigned to papers appearing in different waves, except that graders seemed to need two or three waves to "warm up" in, that is, until they were familiar enough with the overall quality of the papers to designate grades in a consistent manner. Significant differences were found in the quality of writing produced by students at different years in school. The writing of seniors in both content and style was superior to that of sophomores and juniors, the latter two

groups being very similar to one another. Slotnick does note, however, that seniors are not required to take English at Urbana High School, and therefore senior English classes tend to contain the more interested, if not talented, students. Also, no significant differences were found between the average grades assigned by the high school and college raters.

The papers were keypunched and examined by computer using a set of 46 indicators. The computer's measurements were then used in two distinct statistical procedures, a regression analysis and a discriminant analysis, to determine their utility in the prediction of the human quality judgments. Three hundred twenty-six of the papers were used in the two strategies; the remaining 150 were retained for use in cross-validation of the accuracy of the regression and discriminant approaches.

The first prediction procedure involved the step-wise regression of the 46 measures on the grades assigned by the judges. The second procedure divided the papers into five homogeneous groups according to content grades and analyzed then to determine what differences, if any, existed between the groups in terms of the 46 indicators. Papers were assigned the grade associated with the group of papers they most closely resembled. Grade prediction was examined by determining how well the original papers could be assigned to their respective groups on the basis of the discriminant results. The papers were then divided into five groups in terms of style, and a discriminant analysis was again performed. The two grade prediction procedures were cross-validated by predicting the grades of the

remaining 150 papers and comparing them with the grades actually assigned by the human judges. Correlations between the mean grades assigned by the judges and those predicted by the regression and discriminant approaches were significant at .01 level of probability. However, content prediction in this study was achieved more accurately by regression than by discriminant analysis.

Slotnick and Knapp (1971) after briefly reviewing some of the work already done in the area, report that "for now, the computer grading of essays is simply an interesting laboratory phenomenon" since not enough is known about the capability and utility of essay grading by computer to warrant its use as a production tool for English teachers. The purpose of their article is to examine the implications and limitations of that phenomenon. Among implications they list are (1) growth of consciousness of purpose on the part of graders, (2) multiple grading of essays--we can get the computer to predict what any number of professors would have marked any given essay, (3) atypical papers would be spotted (say 1 in 100 or 1 in 500)--either those in dire need of help or budding James Joyces, (4) development of norms based on large-scale exams, in the sense that it would be possible to say, "This is how students of a given age, sex, and/or grade write," (5) student placement--colleges which develop prediction equations can examine writing samples produced by their applicants and make statements about each applicant's probable success as a student writer, (6) the computer could generate first rewrite suggestions--this would relieve the teacher of some theme grading demands while requiring that he retain responsibility for

grade assignments. To facilitate this we need to develop accurate character readers (i.e., a machine which can read typed or printed matter and put it on magnetic tape or punch cards). The authors note that character readers claiming some fair degree of accuracy already exist.

They conclude by stating that the approximate qualities we examine must extend more solidly into the aspect of content. We must determine how intrinsic qualities vary from one age group to another, from topic to topic, mode to mode, etc. Much more must be learned about language and the process of human communication.

Thomas Knapp's paper (1972) reports the result of a study concerned with the effect of essay topic and mode of discourse assigned on some of the mechanical measures which had been found in previous investigations to be predictive of overall quality. Six hundred forty-three secondary school students from the Rochester, New York area, stratified by grade (8th and 11th), sex, and community type (inner city, suburban, rural) were randomly assigned one of nine essay-writing tasks designed to tap three broad topics (self, school, society) at each of three modes of discourse (narrative, descriptive, argumentative). All the essays were written under typical class conditions. Each was coded and keypunched with all errors intact. Knapp found that the topic and mode of discourse did have an effect on the mechanics of essay writing. This effect was strongest on total number of words (the narrative mode was lengthiest for all three topics), total number of common words (most for argumentative--self and narrative--school), number of interrogative sentences (most for argumentative--society and argumen-

tative--school) and standard deviation of word length (greatest deviation for all three modes on topic of society).

More of these types of studies must be undertaken if we are to learn enough about the process of writing and grading to enable us to accurately analyze the product. "In summary, the analysis of student writing seems one of the major educational contributions which the computer is destined to make. Such essay analysis has always been an important job for the teacher, demanding his best dedication and intelligence. Therefore it is not surprising that mechanical 'dedication' and artificial 'intelligence' should assume some of the responsibility in our increasingly computerized world" (page, 1968).

References

Daigon, A. Computer grading of English composition. English Journal, 1966, 55, 46-52.

Hiller, J. H., Marcotte, D. R., & Martin, T. Opinionation, vagueness, and specificity-distinctions: Essay traits measured by computer. American Educational Research Journal, 1969, 6, 271-86.

Knapp, T. R. Essay topics and modes, and their effects on student prose. Paper presented at the Annual Meeting of the American Educational Research Association. Chicago, April, 1972.

Page, E. B. The imminence of grading essays by computer. Phi Delta Kappan, 1966, 47, 238-43.

Page, E. B. Grading essays by computers: Progress report. Proceedings of the 1966 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1967, pp. 87-100.

Page, E. B. The use of the computer in analyzing student essays. International Review of Education, 1968, 14, 210-24.

Slotnick, H. B. An examination of the computer grading of essays. Unpublished doctoral dissertation, University of Illinois, 1971.

Slotnick, H. B. & Knapp, J. V. Essay grading by computer: A laboratory phenomenon?" English Journal, 1971, 60, 75-87.